HTSstation

a user-friendly High-Throughput Sequencing data analysis with evolvable open source modules

Fabrice PA David, Marion Leleu, Julien Delafontaine, Gregory Lefebvre, Frederick Ross, Yohan Jarosz, Lucas Sinclair and Jacques Rougemont.

Bioinformatics and Biostatistics Core Facility
School of Life Sciences
EPFL

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

ISREC — FONDATION | STIFTUNG | FOUNDATION

hosted by Vital-IT — High Performance Computing Center

General Architecture

Web application
for basic usage

http://htsstation.vital-it.ch

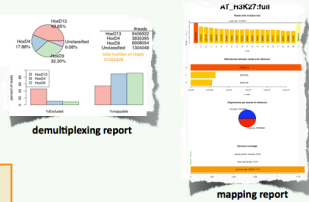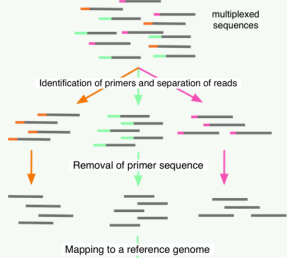### Library & Scripts
for advanced usage

- Analysis modules are implemented in the **bbcflib** python library
- Analyses can be run at command line with **bbcfutils** python scripts accepting more advanced options than on the website
- Pipelines are built and managed with **Bein**
- Codes of Bein, the library and scripts are **open source**

http://github.com/bbcf

① Choose your analysis
② Analyze your data
③ Download your results

Demultiplexing & Mapping

Separate & map to a reference genome

- Separate reads, remove barcodes and filter undesired sequences
- Use bowtie to map reads to genome
- Include, and compensate for, multiple mappings in densities
- Remove multiple identical reads (PCR artifacts)

multiplexed sequences

Identification of primers and separation of reads

Removal of primer sequence

Mapping to a reference genome

demultiplexing report

mapping report

RNA-seq

Transcriptome & exonome profiling

- Calculate counts and RPKM per transcript, gene and exon from a single alignment
- "Optimal" estimation of transcripts expression
- Interactive MA-plot of differential expression between samples

$X = \log2(x1/x2)$

$Y = \log10(\sqrt{x1 \cdot x2})$
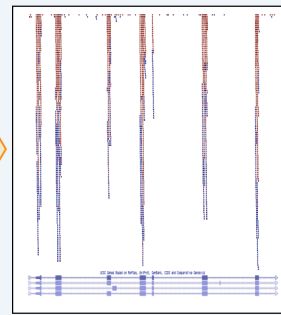
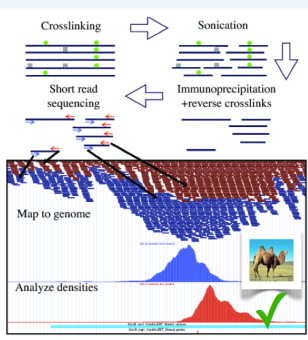Reads from RNA sequencing are aligned on annotated exonome
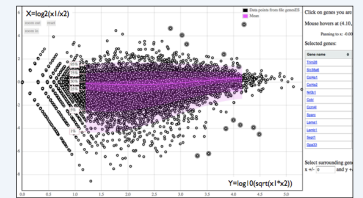
Interactive MA-plot to visualize differential gene/transcript expression

ChIP-seq

Locate & quantify all protein-bound sites

- Profiling interactions between (1) protein and (all) DNA
- Low rate of false positives
- High spatial resolution
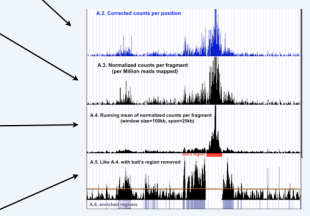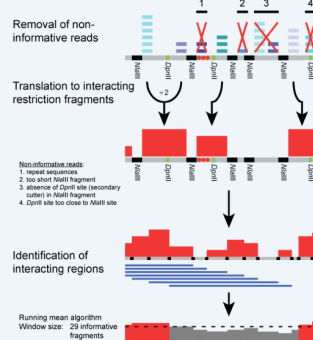- Precise quantitation
- Large dynamic range

Crosslinking — Sonication

Short read sequencing — Immunoprecipitation +reverse crosslinks

Map to genome

Analyze densities

and correct for artifacts ...

PCR duplicate...

repeats...

4C-seq

Identification of long-range 3D-interaction partners

- Create library of fragments
- Identify informative fragments
- Calculate a normalized score per informative fragment

Removal of non-informative reads

Translation to interacting restriction fragments

Non-informative reads:
1. repeat sequences
2. too short NlaIII fragment
3. absence of DpnII site (secondary cutter) in NlaIII fragment
4. DpnII site too close to NlaIII site

Identification of interacting regions

Running mean algorithm
Window size: 29 informative fragments
FDR: 0%

A.2. Corrected counts per position
A.3. Normalized counts per fragment (per Million reads mapped)
A.4. Running mean of normalized counts per fragment (window size=100kb, span=29kb)
A.5. Like A.4 with bait's region removed
A.6. enriched regions

Visualization of 4C-seq results

GDV & gFeatMiner

Manage, Export & Share your data

Validate & Generate new hypothesis
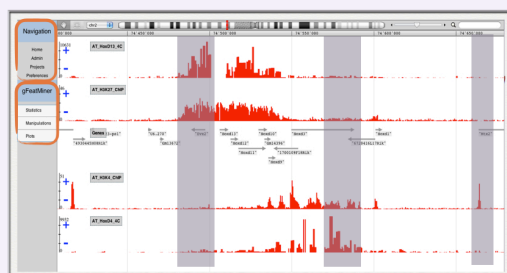
### GDV: Fast, responsive & interactive visualization

- Intuitive navigation
- Share projects between collaborators
- Import genomes and related data from private or public servers
- Customized views
- Select regions of interest and generate new tracks
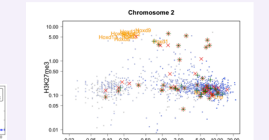
gFeatMiner: Perform interactive analyses

- Descriptive statistics
- Advanced plots
- Data manipulations
- *and more in the future (data mining, search motifs...)*
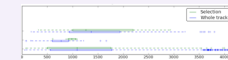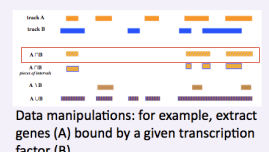
Data manipulations: for example, extract genes (A) bound by a given transcription factor (B)

A ∩ B
A ∖ B
A ∪ B

Descriptive statistics

Advanced plots

Chromosome 2

Summary

- Processing and visualizing the data is a prerequisite to proper biological analysis and interpretation
- Integrate multiple genome-wide datasets for a deeper insight into biological mechanisms
- Modular framework to easily integrate new analysis algorithms
- To be completed with other modules (re-sequencing, downstream analysis,...)

References

- Leleu M, Lefebvre G, Rougemont J. *Processing and analyzing ChIP-seq data: from short reads to regulatory interactions*, Brief Funct Genomics, 9(5-6):466-476, 2010.
- Rey G, Cesbron F, Rougemont J, Reinke H, Brunner M, Naef F. *Genome-Wide and Phase-Specific DNA-Binding Rhythms of BMAL1 Control Circadian Output Functions in Mouse Liver*, PLoS Biol., 9 (2):e1000595, 2011.
- Rougemont J., & Naef F. *Computational Analysis of Protein-DNA Interactions from ChIP-seq Data* Methods in molecular biology, 786:263–273, 2012.
- Noordermeer D, Leleu M, Splinter E, Rougemont J, De Laat W, Duboule D, *The dynamic architecture of Hox gene clusters*, Science, 334:222-225, 2011.
- Gheldof N., Leleu M., et al. *Detecting Long-Range Chromatin Interactions Using the Chromosome Conformation Capture Sequencing (4C-seq) Method* Methods in molecular biology, 786:211–225, 2012.